http://dx.doi.org/10.1590/1678-460X2021370104

DELTA

Articles

French *de* and *en* as expressions of the genitive case: a unified analysis within LFG and computational implementation in XLE[1]

As formas de *e* en *do francês como expressões do caso genitivo: uma análise unificada no quadro da LFG e implementação computacional no XLE*

Leonel Figueiredo de Alencar[2]
Christoph Schwarze[3]

ABSTRACT

The French clitic pro-form en *represents a wide range of heterogeneous constituents:* de-*PP complements and adjuncts, partitive objects, and prepositionless objects of cardinals. The main goal of this paper is to formalize this relationship computationally in terms of genitive case. This is apparently the first non-transformational counterpart to Kayne (1975)'s*

1. We are grateful to Xerox's PARC researchers John Maxwell and Daniel Bobrow for granting this paper's first author a non-commercial XLE license. We thank Valeria de Paiva for acting as intermediary in this process. Thanks are due to Bernard Fradin, Fiammetta Namer, and Florence Villoing for their acceptability judgments. We are also indebted to Harro Stammerjohann, Jessé de Sousa Mourão, and the two anonymous reviewers for helpful comments and suggestions on earlier drafts of this paper, although any remaining errors are our own.

2. Universidade Federal do Ceará. Ceará – Brasil. https://orcid.org/0000-0001-8148-6994. E-mail: leonel.figueiredo.de.alencar@gmail.com.

3. Professor Emeritus, Department of Linguistics. University of Konstanz. Germany. http://orcid.org/0000-0002-6103-9721. E-mail: christoph.schwarze@uni-konstanz.de.

unified analysis, which derives en *from a deep structure with* de *by means of syntactic transformations. Transformational grammars are problematic from the parsing perspective. In order to test our analysis automatically on a large amount of data, we implemented it in a computational grammar of French in the Lexical-Functional Grammar (LFG) formalism using the XLE system. This non-transformational framework is particularly fit for expressing systematic relationships between heterogeneous structures and has successfully been used for the implementation of natural language grammars since the 1980s. We tested the implementation on 320 grammatical sentences and on an equal number of ungrammatical examples. It analyzed all grammatical examples and blocked almost 95% of the ungrammatical ones, showing a high empirical adequacy of the grammar.*

Keywords: *genitive case; prepositions; pronominal clitics; computational linguistics.*

RESUMO

A pró-forma clítica en *do Francês representa ampla variedade de constituintes heterogêneos: PPs complementos e adjuntos introduzidos por* de*, objetos partitivos e objetos desprovidos de preposição de numerais cardinais. O objetivo principal deste artigo é formalizar essa relação computacionalmente por meio do caso genitivo. Esta é, aparentemente, a primeira contraparte não-transformacional da análise unificada de Kayne (1975), a qual deriva* en *de uma estrutura profunda com* de *por meio de transformações sintáticas. Gramáticas transformacionais são problemáticas sob a perspectiva da análise sintática automática. A fim de testar nossa análise automaticamente em um grande volume de dados, implementamo-la em uma gramática computacional do francês no formalismo da Gramática Léxico-Funcional (LFG) usando o sistema XLE. Esse modelo não-transformacional é especialmente adequado para expressar relações sistemáticas entre estruturas heterogêneas e tem sido usado com sucesso na implementação de gramáticas de línguas naturais desde os anos de 1980. Testamos a implementação em 320 sentenças gramaticais e em igual número de exemplos agramaticais. Foram analisados todos os exemplos gramaticais e bloqueados quase 95% dos agramaticais, mostrando que a gramática possui uma alta adequação empírica.*

Palavras-chave: *caso genitivo; preposições; clíticos pronominais; linguística computacional.*

1. Introduction

There is a striking parallelism in French between forms *de* and *en,* see (1)-(8), where the constituent containing *de* in (a) is anaphorically substituted for by the pronominal clitic *en* in (b).

(1) a. La population dépend [de la forêt].[4]
the population depends DE the:F.SG forest
'The population depends on the forest.'
b. La population en=dépend.
the population EN=depends
'The population depends on it.'

(2) a. Luc vient [de Paris].
Luc comes DE Paris
'Luc comes from Paris.'
b. Marie en=vient aussi.
Marie EN=comes too.
'Marie comes from there too.'

(3) a. Il=est fier [de la victoire].
he=is proud DE the:F.SG victory
'He is proud of the victory.'
b. Il=en=est fier.
he=EN=is proud
'He is proud of it.'

(4) a. Elle=a mangé la moitié [de la tarte].
she=has eaten the half DE the:F.SG tart
'She ate half of the tart.'
b. Elle=en=a mangé la moitié.
she=EN=has eaten the half
'She ate half of it.'

(5) a. Marie a acheté trois [de ces pommes].
Marie has bought three DE these apples
'Marie bought three of these apples.'
b. Marie en=a acheté trois.
Marie EN=has bought three
'Marie bought three.'

4. Unless otherwise stated, all examples, glosses, and translations are our own.

(6) a. Les causes de sa maladie sont inconnues.
the causes DE his illness are unknown
'The causes of his illness are unknown.'
b. Les causes en=sont inconnues.
the causes EN=are unknown
'Its causes are unknown.'

(7) a. Jeanne a la clef [du coffre].
Jeanne has the key DE.the.M.SG trunk
'Jeanne has the key of the trunk.'
b. Jeanne en=a la clef.
Jeanne EN=has the key
'Jeanne has the key of it.'

(8) a. Elle=doit acheter [de la farine].
she=must buy DE the:F.SG flour
'She needs to buy flour.'
b. Elle=doit en=acheter.
she=must EN=buy
'She needs to buy some.'

This article pursues two goals. First, we propose a formal account of the relationship between these two forms within Lexical-Functional Grammar (henceforth LFG), a framework particularly fit for expressing systematic relationships between heterogeneous structures (Bresnan, 2001). Its adequacy for implementing computational grammars of natural languages has been continually demonstrated for over the past 35 years (cf. Müller, 2018, p. 219-220). Second, we implement the proposed analysis computationally in the Xerox Linguistic Environment (XLE)[5] as an extension of FrGramm, an LFG grammar fragment of French developed in this system (Schwarze & Alencar, 2016; Alencar, 2017). A computational implementation enables us to check automatically a particular approach to a grammatical phenomenon for empirical validity on a large amount of data.

To our knowledge, our proposal is the first non-transformational unified analysis of *de* and *en*. It distinguishes itself from previous LFG approaches in a two-fold way. First, it explains *en*-pronominalization of a wide range of heterogeneous constituents in terms of a single

5. http://ling.uni-konstanz.de/pages/xle/doc/xle_toc.html

common feature, namely, genitive case. Second, it postulates a single representation for both items. It is a lexicalist counterpart to Kayne (1975)'s transformational analysis, which uniformly relates diverse uses of *en* to a single deep structure representation with the preposition *de*.

The next section presents the basic facts to be modeled. Section 3 then outlines the theoretical framework. After a review of previous directly related approaches in Section 4, Section 5 details the formalization of our analysis. Section 6 deals with the implementation methodology and evaluation results. In the last section we summarize the main conclusions and point out directions for further research.

2. A closer look at the relationship between *de* and *en*

De is a highly ambiguous form. In (1)-(5), it satisfies a subcategorization requirement of a verbal, adjectival, nominal, and numeral head, while in (6) and (7) it introduces an adjunct to a noun in subject and object position, respectively. In (8), however, it does not function as an independent syntactic word, but is instead an element of the multiword determiner *de la* in the partitive direct object. Diachronically, French partitive determiners *du* and *de la* derive from the preposition *de* and the singular masculine and feminine definite article, respectively (Carlier *et al.*, 2013). The status of *de* in constructions of the type of (2) is controversial. For Frank (1996, p. 165), it is a semantic preposition. However, Carlier *et al*. (2013) show it to have semantically bleached (see Section 4). We follow this view here, treating it as a genitive marker in all constructions (1)-(7).

Other usages disallow pronominalization by *en*, e.g., (9)-(11). In (9), *de* is a semantic preposition heading a locative adjunct, while in (10) it introduces an infinitival complement pronominalizable by the accusative clitic *le* (Carlier *et al*., 2013). In (11), it heads a classificative PP (Fábregas, 2017), which does not denote an event participant capable of being pronominalized.

(9) De la digue […] on=aperçoit la pointe […] (Google)
from the seawall one=perceives the headland

(10) Elle=décide de les=en=faire paraître fiers.
she=decides DE them.ACC=EN=make.INF seem.INF proud:M.PL
'She decides to make them seem proud of it.'

(11) Elle=prend le train de Paris.
she=takes the train DE Paris
'She takes the Paris train.'

The systematic relationship between *de* and *en* can be described in terms of the shared functional properties in Table 1.

Table 1 – Functional properties of *de* and *en*

Nº	Function	Examples
(i)	Oblique complement of a verb or adjective	(1)-(3)
(ii)	Domain of a quantifying form	(4) and (5)
(iii)	Adjunct to a noun	(6) and (7)
(iv)	Partitive direct object	(8)

Properties (i), (iii), and (iv) use familiar terminology. Property (ii), however, demands a detailed explanation. Quantified terms are expressions made up of a quantifying form (henceforth QForm) and its domain, e.g., *five apples* or *a liter of milk*. The QForm types in French that were implemented in our grammar are presented in Table 2.

Table 2 – French QForm types implemented in the grammar

Nº	Type	Example
(i)	Cardinal numerals	*trois* 'three'
(ii)	Measure names	*kilo* 'kilo'
(iii)	Fraction names	*moitié* 'half'
(iv)	Collective numerals	*douzaine* 'dozen'

The canonical domain of a QForm is a determiner phrase (DP) or a prepositional phrase headed by *de* (henceforth *de*-PP). If the QForm is a cardinal up to 999999, the choice depends on the set designated by the domain. If it is determined, the domain is a *de*-PP, otherwise it is a DP, see (5) and (12), respectively (Milner, 1978 *apud* Hulk, 1983, p. 168).

(12) Marie a acheté trois pommes.
Marie has bought three apples

With all other QForms of Table 2 the domain is a *de*-PP, see (13)-(16):

(13) deux millions de personnes
two million DE people
'two million people'
(14) deux bons kilos de dynamite (Google)
two good kilos DE dynamite
'two good kilos of dynamite'
(15) une bonne douzaine de correspondants (Google)
a good dozen DE correspondents
'a good dozen of correspondents'
(16) la moitié pauvre des habitants du monde (Google)
the half poor DE.the.PL inhabitants DE.the.M.SG world
'the poor half of the world's inhabitants'

In direct object position, the domain can unconstrainedly be referred to by *en*. The pro-form is obligatory if the QForm is a cardinal, compare (17) with the pronominalized version of both (5a) and (12) in (5b). By contrast, the grammaticality of *en* as an OBJ of a QForm in preverbal subject position has generally been denied (Hulk, 1983; Kayne, 1975; Lagae, 1997), but positive evidence can be found with Google, e.g., (18), extracted from a French archeology journal. Although three native-speaker informants were unhappy with this example, we hypothesize that this configuration is optionally licensed for passive or unaccusative verbs and copulas, at least in formal language.

(17) *Marie a acheté trois.
Marie has bought three

(18) […] des tombes ont été aménagées sur ce site.
ART.INDF.PL graves have been installed on this site
'Graves were installed at this site.'
Deux en=ont été fouillées en 1980 […].
two EN=have been excavated in 1980
'Two of them were excavated in 1980.'

3. The theoretical framework: Lexical-Functional Grammar

Lexical-Functional Grammar (LFG) is a non-transformational generative model that strictly adheres to the Lexical Integrity Principle, only allowing transformations in the lexicon (Bresnan, 2001). It factors the syntactic analysis of a sentence into two distinct representation levels related by a projection function: f(unctional)-structure and c(onstituent)-structure. These are exemplified in **Figure 1** and **Figure 2**, respectively.

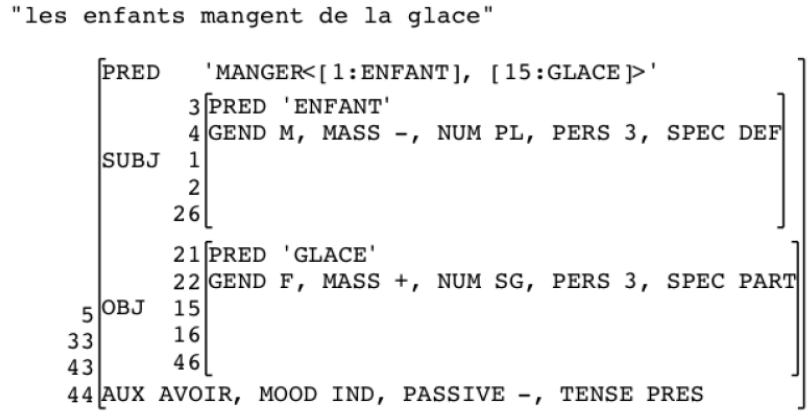


Figure 1 – *F-structure for example (19).*

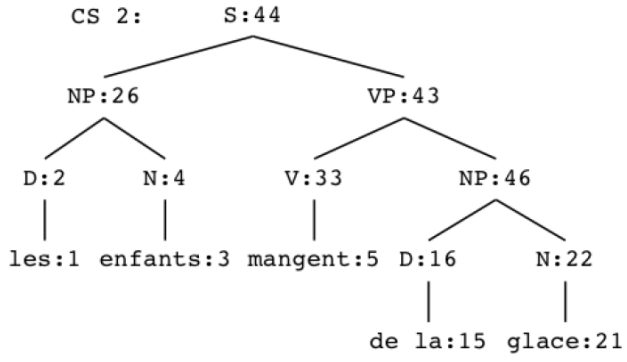


Figure 2 – *C-structure for example (19).*

(19) Les enfants mangent de la glace.
the children eat DE the:F.SG ice_cream
'The children eat ice cream.'

These representations were automatically generated by XLE from the grammar in (20)-(26). Numerical indexing indicates the mapping between f-structure and c-structure, i.e., each c-structure node in **Figure 2** is labeled with a number that labels the corresponding f-structure in **Figure 1.**

C-structure expresses precedence and part-whole relations between the constituents of a sentence. For example, the representation in **Figure 2** shows that the sentence (S) is made up of a noun phrase (NP) and a verb phrase (VP). The latter phrase, in turn, consists of a verb (V) and an NP made up of a determiner (D) and a noun (N).

F-structure encodes morphosyntactic aspects like agreement, tense, subcategorization, etc. in a format of features. A feature consists of an attribute and a value, e.g., GEND(er)=F, NUM(ber)=SG, and SPEC(ification)=PART(itive) in f-structure 46 of **Figure 1** state that *de la glace* is feminine, singular, and partitive.[6] An attribute must not have divergent values, e.g., NUM=SG and NUM=PL(ural) in f-structure 1, if we substitute *l'enfant* 'the child' for *les enfants* 'the children' in (19). A value may be atomic, e.g., PL, or constitute a feature structure, as is the case with the value of grammatical functions, e.g., SUBJ (subject) and OBJ (object). The value of a PRED(icate) attribute is a semantic form, which includes the subcategorization frame of valence-bearing lexemes enclosed in angle brackets.

The syntactic components of an LFG grammar are constituency rules and lexicon entries, exemplified in (20) and (21)-(25), respectively.

6. In the f-structures throughout the paper, category names with up to 5 characters (and a few longer names) are not abbreviated. Abbreviations for categories dealt with in the main text or in footnotes are explained as introduced. Otherwise, the following abbreviations are used: ACC=accusative, ATTRIB=attributive, ATYPE=adjective type, AUX=auxiliary, CFORM=complementizer form, DCONTR=preposition and determiner contraction, DECLAR=declarative , DEF=definite , DEM=demonstrative, DTYPE=determiner type, IND=indicative, INDEF=indefinite, INF=infinitive, M=masculine, NEG=negation, NEGP=negation particle, NOM=nominative, NOSEM=nonsemantic, PART_PAST=past participle, PERS=person, POSTNOM=postnominal, PTYPE=preposition type, PREDIC=predicative, PRES=present, UNACC=unaccusative, VFORM=verb form.

For the sake of readability, we adapted XLE's syntax, as far as possible, to the traditional LFG notation. Throughout the paper, the ellipsis indicates omission of code irrelevant to the discussion, e.g., (23). Both constituency rules and lexicon entries are endowed with functional annotations, where "↓" refers to the feature structure of the node the annotation is attached to, while "↑" denotes the feature structure of its mother node. These annotations indicate the f-structures the c-structure nodes project to. A semicolon separates the annotations pertaining to a node from those of its sister category.[7] Functional annotations mostly have the form of equations in the form (f ATTRIBUTE)=VALUE, assigning VALUE to ATTRIBUTE of f, where f is a feature structure. For example, (↑MASS)=+ in (24) assigns MASS=+ to D, blocking an NP with a MASS=– (i.e., count) noun head.

(20) S → NP: (↑SUBJ)=↓; VP.[8]
VP → V {NP:(↑OBJ)=↓|PP:(↑OBL)=↓}#0#1.
NP → D N.
PP → P NP.

(21) *les*, D
(↑SPEC)=DEF
(↑NUM)=PL
(↑PERS)=3

(22) *enfants*, N
(↑PRED)='ENFANT'
(↑GEND)=M
(↑NUM)=PL
…

(23) *mangent*, V
(↑PRED)='MANGER<(↑SUBJ)(↑OBJ)>'
(↑SUBJ NUM)=PL
(↑SUBJ PERS)=3
(↑TENSE)=PRES
(↑MOOD)=IND
...

7. In the traditional notation, annotations are written below the respective nodes, demanding much more space.
8. In XLE, ↑=↓, which identifies the f-structures of daughter and mother modes, is automatically attached to nodes deprived of further equations.

(24) *de la*, D
(↑NUM)=SG
(↑GEND)=F
(↑PERS)=3
(↑SPEC)=PART
(↑MASS)=+

(25) *glace*, N
(↑PRED)='GLACE'
(↑GEND)=F
(↑NUM)=SG

The rules in (20) define the c-structure and f-structure of S, VP, NP, and PP (prepositional phrase), using the categories D, V, P, and N from the lexicon. An annotation of the form (↑GF)=↓, where GF is a grammatical function, states that the constituent in question realizes function GF of the mother category. Thus, in the first rule, NP is the SUBJ of S. In the second rule, NP is the OBJ and PP is the oblique (OBL) of VP. These two categories are connected by the Boolean disjunction "|", marked as optional with #0#1[9], thereby licensing VPs without any complements.

The completeness and coherence conditions ensure that all and only the governed grammatical functions listed in a predicate subcategorization frame are realized in the syntax. For example, the verb *manger* 'eat' requires a SUBJ and an OBJ, see (23).

Provided with additional entries, e.g., (26) and (27)[10], this small grammar fragment is also capable of generating (1a), producing the representations in **Figure 3** and **Figure 4**.

```
"la population dépend de la forêt"

      [PRED  'DÉPENDRE<[1:POPULATION], [9:FORÊT]>'
      SUBJ 1[PRED 'POPULATION']
    5 OBL  9[PRED 'FORÊT']                          ]
```

Figure 3 – *Simplified f-structure for example (1a).*

9. XP#m#n means from m to n repetitions of XP.
10. Both based on Schwarze (1996), see Section 4.2.

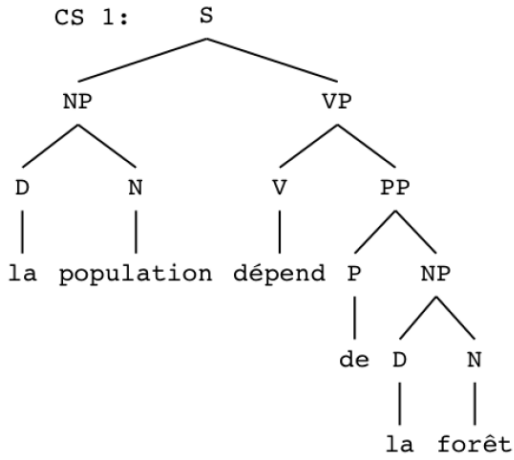


Figure 4 – *Simplified c-structure for example (1a).*

(26) *dépend*, V
(↑PRED)='DÉPENDRE<(↑SUBJ)(↑OBL)>'
(↑OBL PCASE)=c DE
@(V-INFL @PRS @IND 3 SG)

(27) *de*, P
(↑PCASE)=DE

Entry (26) states that verb *dépendre* 'depend' subcategorizes for a SUBJ and an OBL, the prepositional case (PCASE) of which must be DE. This requirement is satisfied by entry (27). Notation "=c" in (26) represents a *constraining* equation. While defining equations with "=" set the value of an attribute, constraining equations require its value to be defined elsewhere, in the case at hand, by the lexical entry of the preposition.

The inflectional features of the verb are encoded by means of a *template* in the last line of (26). In XLE, templates are analogous to functions in a programming language, enabling code reuse, so that the same blocks of commands need not be written over and over again. Template definitions have the general form *NAME=definition* or *NAME(P1 P2… Pn)=definition*, where P1, P2, etc. are parameters. According to the definition in (28), V-INFL takes four parameters: T(ense), M(ood), P(erson), and N(umber). "@" is the template call

operator, which has the following syntax: *@NAME* or *@(NAME P1 P2… Pn)*. In the call to V-INFL in (26), the first two parameters are themselves template calls, see definitions in (29). When processing a grammar, XLE substitutes definitions for template calls, instantiating the parameters.

(28) V-INFL(T M P N)=T M
(↑SUBJ NUM)=P
(↑SUBJ PERS)=N

(29) PRS=(↑TENSE)=PRES
IND=(↑MOOD)=IND

4. Previous approaches

There is a vast literature on Romance pronominal clitics, within different theoretical frameworks, e.g., Heap *et al*. (2017). In generative grammar, it seems that much less attention has been paid to grammatical prepositions such as *de*, despite the overall awareness that *en* is related to *de*-PPs. We limit ourselves here to what is immediately relevant to our own analysis.

Due to the historical connections between the different uses of *de* and *en*, we first summarize the study by Carlier *et al*. (2013) on the grammaticalization of these two elements from the Classical Latin period onwards. In the next subsection we mention some studies carried out in the framework of early generative grammar, before dealing with LFG analyses in 4.2.

According to Carlier *et al*. (2013), the Latin ablative preposition *de*, whose core meaning was spatial distancing from a source, underwent extension to other domains in the course of time, e.g., origin and lineage, extraction, partition, and inclusion. Two additional parallel developments led to the distribution of *de* and *en* in Modern French. First, the Latin genitive case, whose core function was linking two NPs in a possession relation, was progressively substituted for by *de*-PPs. Second, the pronoun *en*, derived from the Latin adverb *inde* 'from there', underwent successive bleaching in Medieval French and spread as a replacement of *de*-PPs with different non-spatial meanings, e.g., as the domain of a cardinal.

Carlier *et al*. (2013) show that the continuation of these developments produced a threefold result in Modern French. First, *de* lost semantic content. This, in turn, brought about major changes in its usage. For example, it may introduce the complement of verbs with opposite meanings such as *s'approcher de* 'to come closer to' and *s'éloigner de* 'to get further from' and is often used with other elements to reinforce the spatial meaning of a verb complement, see (30). While *de* retains its status as a semantic preposition in locative adjuncts, typically with perception verbs as in (9), it is reduced to a genitive marker of noun or verb arguments, see (1a). Second, the combination of *de* and definite article grammaticalized into a full-fledged partitive article, see (8a). Third, *en* fully desemantized and became a clitic pro-form for genitive objects and quantified direct objects (i.e., partitive objects), see (1b) and (8b).

(30) Elle revien-t du médecin / de chez le médecin
she come-PRS-3SG of.the doctor / of at the doctor
'She comes from the doctor' (Carlier *et al*., 2013, p. 43, their translation and glosses)

Two critical remarks to Carlier *et al*. (2013) are in order. First, they do not deal with *de*-PPs functioning as adjective complements, domain of a cardinal, or adjunct to a noun. Second, they treat genitive objects and quantified direct objects as two separate functions without any common link. We address these issues in our unified account in Section 5.

4.1. Earlier generative research

Kayne (1975, p. 107-110) uniformly categorizes *en* as a *de*-PP pro-form even in cases where it does not correspond to an overt PP as in (12). He argues that this pro-form is derived from a deep structure representation with *de* by means of syntactic transformations.

Hulk (1983) opposes Kayne's unitary solution. For the quantitative construction in (12), she proposes an additional PRO-N' variant, i.e., a pro-form for the intermediary projection N-bar. She argues that this type of quantitative NP is derived from (31), where the Spec(ifier) is

marked with +Q, i.e., it is a quantitative determiner. The corresponding constructions with *en* result from pronominalization of N', e.g., (5b).

(31) $_{NP}[_{Spec}$ [+Q] $_{N'}$[de N]]

According to Hulk, examples like (5b) are ambiguous, since they also have a "partitive" interpretation, corresponding to (5a). For these "partitive" NPs, she proposes (32), where α represents an empty N head. *En* pronominalizes the *de*-PP in this construction.

(32) $_{NP}[_{Spec}$ [+Q] $_{N}$[α] $_{PP}$[de NP]]

Hulk motivates the distinction between (31) and (32), among other evidence, with agreement facts, cf. (33) and (34), respectively. (33) is ungrammatical because Spec and head N do not agree in number. In (34), by contrast, Spec and head N need not agree, since an empty N is not marked for number. Note, however, that these two constituents must agree in gender, so that Hulk's analysis fails to predict the ungrammaticality of (35).

(33) *un livres
one:M.SG book(M):PL
(34) un de ces livres
one:M.SG DE these book(M):PL
'one of these books'
(35) *une de ces livres
one:F.SG DE these book(M):PL

In the 1990s, generative linguistics abandoned the transformational frameworks underlying these two approaches. As Klenk (2003, p. 78-80) shows, parsing with transformational grammars is difficult, if not impossible, especially in case of deletion transformations. Kayne's analysis, however, is still inspiring, in that it tries to capture the systematic relationship between *en* and *de*-PPs in a unified way.

In Section 5, we propose a unified lexicalist account of this relationship without resorting to transformations, while at the same time handling the agreement facts in (33)-(35) and also examples like (18), considered ungrammatical by Kayne and Hulk.

Jones (1996) categorizes *en* as a pro-PP and proposes that the function of grammatical prepositions is to assign Case to an NP. This approach eliminates the need for deriving *en* from a deep structure with *de*, preparing the ground for a unified analysis in terms of case. However, Jones (1996) did not undertake such analysis, which we do in Section 5.

4.2. Previous LFG analyses

The grammar of French clitics has been a topic in LFG since the origins of the model. Grimshaw (1982) treats them as members of the clitic category (CL) expanding V to V'. The lexical entries proposed comprise case features. However, *en* and *y* are disregarded.

Schwarze (1996) was one of the first lexical-functional accounts of the systematic relationship between *en* and *de*-PPs. He argues that the nonsemantic *de* has the same function as the genitive suffix in languages with morphological case like German. For constructions (1a) or (3a), he provides *de* with the feature *(↑PCASE)=DE* and assigns the same feature to *en*, see (27) and (36). Accordingly, entries for the corresponding predicators must contain a constraining equation requiring the oblique to have PCASE=DE, see (26), an adaptation of Schwarze's partial entry for *parler* 'to speak'. The proposed analysis of *de*-PPs, but not of *en*, was tested on an LFG parser.

(36) *en*, CL, (↑PRED)='PRO', (↑SPEC)=DEF, (↑PCASE)=DE

This approach, however, has some drawbacks. First, the notion of PCASE is inappropriate within the system of French pronominal clitics: subject and direct object clitics correspond to noun phrases without a PCASE, so that it seems more reasonable to establish distinctions based on traditional morphological cases, e.g., Heap *et al*. (2017, p. 189-193). Second, it treats *en* as three-way ambiguous, proposing two additional variants with *(↑SPEC)=PARTITIV* that only differ from one another in the grammatical function they perform, namely direct object and MOD(ifier) of a direct object. As we will show, such lexical ambiguity can be avoided. Third, the corresponding constructions with *de*, e.g.,

(8a) and (4a), are not taken into account, nor is the use of both *de* and *en* as adnominal adjuncts in examples like (6) and (7).

Frank (1996) is so far the most complete documentation on a large-coverage grammar of French in the LFG formalism.[11] Lexicon entries of key items and all syntactic rules are described in detail. The complete source code of the syntactic component is made available in an appendix.

Although Frank's analysis has a much wider scope than Schwarze (1996)'s, both share strong similarities in the treatment of the common subset of constructions with *de* and *en*. Remarkably, however, these two coetaneous studies do not cite each other. Frank proposes that heads governing obliques with the nonsemantic *de*, as in (1a) and (3a), subcategorize for a DE OBJ, where the attribute DE is provided by the PCASE value of the preposition. The clitic *en* is likewise assigned the feature PCASE=DE, so that the corresponding constructions with *en* can be analyzed, cf. (1b) and (3b).

Complex annotated c-structure rules handle a large range of different possibilities of clitic placement and of internal arrangement of clitic clusters, undoubtedly one of the most intricate aspects of French syntax, due to the variety of intervening factors. One of the complexities that are successfully dealt with is pronominal clitic climbing, e.g., (5b), where the pronoun moves from its canonical proclitic position in relation to its governing head, as in (8b), to the left-adjacency of the auxiliary. Causative *faire* 'make' and the copula behave similarly as far as pronoun climbing is concerned: pronominal clitics are attracted to the leftmost member in a series of verbs of this type, cf. (10) and (37)-(39) from our handcrafted corpus. Negation introduces a further complexity layer when combined with pronominal clitics: negative particles (e.g., *ne* and *pas*) also cliticize to the verb, either enclosing an individual pronoun or a pronominal cluster and its finite host or preceding both pronominal clitics and their infinitive host, cf. (37)-(39).

(37) Marie ne=les=en=a=pas fait parler.
Mary NEG=them.ACC=EN=has=NEG made speak.INF
'Mary didn't make them speak of it.'

11. See Müller (2018, p. 219) for a comprehensive list of LFG implementations of French. Alencar (2017) compares the coverage of some of these grammar fragments.

(38) Elle=veut ne=pas=l'en=faire parler.
she=wants NEG=NEG=3ps.ACC=EN=make.INF speak.INF
'She wants not to make him speak of it.'

(39) Elle=veut ne=pas=en=être remerciée.
she=wants NEG=NEG=EN=be thank:PTCP.F.SG
'She wants not to be thanked for it.'

In LFG, control verbs subcategorize for an XCOMP, a predicative complement whose open SUBJ slot is filled by a grammatical function of the matrix clause (cf. Bresnan, 2001). In Frank's grammar, auxiliaries are control verbs, alongside copulae, causatives, modals, and aspectual verbs. Clitic climbing is licensed by control verbs lacking a complementizer, i.e., copulae, auxiliaries, and causatives. By contrast, control verbs with a complementizer, e.g., *COMPL FORM=de* in case of *décider* in (10), disallow climbing. Modals are assigned *COMPL FORM=null*, so that they are also unable to host climbed clitics, cf. (8b).

Frank proposes different constituency rules for generating clitic clusters with negation and up to two pronouns in the different varieties of finite and infinitive structures. Due to space limitations, the full details of the implementation cannot be presented here; we focus on the sentence type exemplified in (37). Disregarding the functional annotations for now, the complex formed by a clitic cluster and a finite verb in this construction is generated with the rules in (40)-(43), where the brackets-enclosed constituents are optional. The IP category consists of the I2 complex, formed by a finite verb and (optional) clitics, and zero or more complements. NEGAT and NEGP introduce negation *ne* and negative particle (e.g., *pas*), respectively. Analogous rules are proposed for the other types of structures containing verb clitics.

(40) IP → I2 (NP) (PP)...
(41) I2 → (NEGAT) I1 (NEGP)
(42) I1 → (CL) V
(43) CL → CL1 (CL2)

In the lexicon, pronominal clitics are classified as CL1 and/or CL2. These categories receive appropriate annotations to constrain the types of clitics capable of occupying the respective positions. Since *en* can occur both alone and as the rightmost element in a pronominal cluster,

it is assigned two entries with identical annotations, the only difference being the category label, cf. (44).

(44) *en*, CL1, (↑PRED)='PRO', (↑FORM)=pro, (↑PCASE)=de

For CL1 and CL2 in examples like (1b) and (37), Frank proposes the rule annotation in (45), which contains a disjunction with two alternatives, corresponding to local and non-local cliticization. The equation in the first disjunct states that the clitic is the DE OBJ of V, which is the case with (1b), while the equation in the second disjunct states that the clitic is the DE OBJ of an embedded VCOMP (a verbal XCOMP in her terminology), as in (37), where *en* is a complement of *parler* 'speak'. The other element of the second disjunct is a negative existential constraint specifying that the VCOMP governing the clitic have no COMPL ("¬" symbolizes negation). VCOMP+, where the plus sign symbolizes one or more instances of the preceding string, means that the VCOMP whose verb governs the clitic can be embedded in another VCOMP (which, in turn, can be nested in another VCOMP, and so on).

(45) {(↑DE OBJ)=↓|(↑VCOMP+ DE OBJ)=↓¬(VCOMP+ COMPL)}

Frank's grammar accounts for the parallel behavior of *de* and *en* in only two constructions of (1)-(8), i.e., as the oblique of a verb or adjective, cf. (1) and (3). The use of *en* in the other six construction types was not implemented. Examples like (2b) cannot be analyzed because, according to Frank, this verb type subcategorizes for a thematic oblique PP with *PCASE=source*, which is incompatible with the proposed entries and c-structure rules, cf. (40)-(45).

In constructions of the type in (4a), the PP is analyzed as a "partitive object" (PART OBJ), for which case a variant of *de* with *(↑PCASE)=part* is postulated. Only types (ii) and (iv) of Table 2 were implemented, i.e., measure names and collective numerals. According to Frank, the former require singular mass objects, while the latter require plural non-mass nouns. Again, there is no corresponding variant of *en*.

As regards (5), she categorizes cardinals as NUM. They are optionally generated in an NPDET projection between an optional

DET (i.e., article) position and an obligatory NPMOD projection, which comprises head noun and modifiers, see (46)-(50). A NUM like *trois* 'three' does not subcategorize for a complement. Instead, it just provides a *VALUE=trois* feature to the NP. While examples like (12) can thus be analyzed with Frank's grammar, examples like (5a) and (5b) cannot, since both constructions lack the noun head required under NPMOD, see (47)-(50).

(46) NPDET → (DET) (NUM) NPMOD
(47) NPMOD → …(AP) NMOD A* …NPKOMPL
(48) NPKOMPL → (PP)…
(49) NMOD → NK
(50) NK → N (A) (PP)

Genitive PPs such as (6a) and (7a) are generated as adjuncts under NPKOMPL (i.e., noun complements), see (47)-(48), and assigned the feature *ROLE=obl_poss*, provided by an additional semantic variant of *de*. Partitive determiners, see (8a), are encoded in the lexicon as indefinite determiners, an equation of the form *(↑CLASS)=c mass* ensures that they are only combined with mass nouns. Since Frank's grammar has no corresponding variants of *en*, it cannot analyze (6b)-(8b).

Butt *et al*. (1999) report on the development of large-coverage parallel LFG grammars for English, French, and German. In the French grammar, the clitic *y* is assigned an f-structure with a *PCASE=À* feature similar to that of an adverbial *à*-PP. However, the source code is not publicly available and the implementation details are sparse, without information on how (if at all) examples like (1)-(8) are analyzed by the French grammar.

For Schwarze (2001), *en* is ambiguous between two functions: "Oblique" in cases like (1b) and "Partitive Modificator of the DIRECT OBJECT" in cases like (4b). By contrast, Schwarze (2012) assigns *en* three case values: GEN(itive), ABL(lative), and PART(itive). The question of how the mapping of case features onto grammatical functions actually comes about is left open.

In sum, Frank (1996) is the most complete LFG implementation available of the systematic relationship between *de* and *en*, yet it covers

only a small subset of the data in (1)-(8), which are accounted for by the proposed unified analysis, presented in the following section. Another issue deserving improvement in Frank's implementation is the proliferation of entries for *en* and *de*. Schwarze (2012) is a precursor to our present account, insofar as it abandons the PCASE-based analysis of *en* in favor of traditional case distinctions. However, it is little formalized and treats *en* as three-way ambiguous. By contrast, we propose an implemented (and thereby completely formalized) grammar fragment, which enables us to automatically test it on a large amount of data. This grammar has just one lexical representation for the clitic *en*, besides being able to handle the full range of uses in (1)-(8).

5. A unified account

In this section, we first show how the grammatical preposition *de*, clitic *en*, and partitive determiners should be represented in the lexicon in order to account for their systematic relationship. Then, in 5.1-5.3, we detail our analysis of properties (i)-(iii) in Table 1, exemplified by structures generated by the parser. We focus here on the c-structure rules for *de*-PPs, postponing corresponding rules for *en* to Section 5.4.

The basic idea of our approach is to assume a CASE=GEN(ITIVE) feature as a means to account for the functional relationship between *de* and *en*:

(51) *en*, CL
(↑PRED)='PRO'
(↑CASE)=GEN

(52) *de*, P
(↑CASE)=GEN
…

In view of this analysis, entry (24) for the partitive determiner *de la* is unsatisfactory. Since it does not share any feature with (51), *en*-pronominalization seems fortuitous. To solve this problem, two options are available: (a) replacing the third line of (51) with {(↑CASE)=GEN|(↑SPEC)=PART}, stating that either the case is

genitive or the specification is partitive, or (b) appending a genitive case feature to (24), whereby we obtain (53).

(53) *de la*, D
…
(↑CASE)=GEN

Alternative (a) has an undesirable side effect: it creates a potential source of parsing ambiguity, since each disjunct represents a different lexical variant of *en*. On the other hand, alternative (b) assumes a single lexical representation for *en*, so we consider it preferable.

Figure 5, Figure 6, Figure 7, and **Figure 8** exemplify the parsing of DP and *en* partitive OBJs.

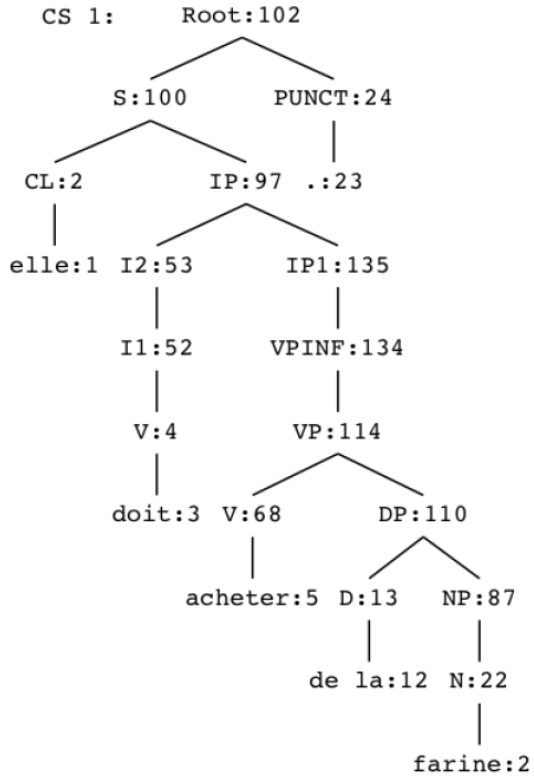


Figure 5 – *C-structure for example (8a).*

DELTA
37.1
2021

```
"Elle doit acheter de la farine."

      [PRED     'DEVOIR<[1:PRO], [5:ACHETER]>'                                                    ]
      [SUBJ     1[PRED 'PRO'                                                                      ]
                2[CASE NOM, GEND F, NUM SG, PERS 3]                                               ]
                [PRED   'ACHETER<[1:PRO], [12:FARINE]>'                                           ]
                [SUBJ   [1:PRO]                                                                   ]
   23           [     21[PRED 'FARINE'                                                            ]
   24           [     22[CASE GEN, DCONTR -, GEND F, MASS +, NUM SG, PERS 3, SPEC PART            ]
    3 [XCOMP     5[OBJ  87                                                                        ]
    4           [       12                                                                        ]
   52          68       13                                                                        ]
   53         114      110]                                                                       ]
   97         134]                                                                                ]
  100         135[AUX AVOIR, CFORM null, PASSIVE -, VFORM INF                                     ]
  102[CLAUSE_TYPE DECLAR, MOOD IND, PASSIVE -, TENSE PRES                                         ]
```

Figure 6 – *F-structure for example (8a).*

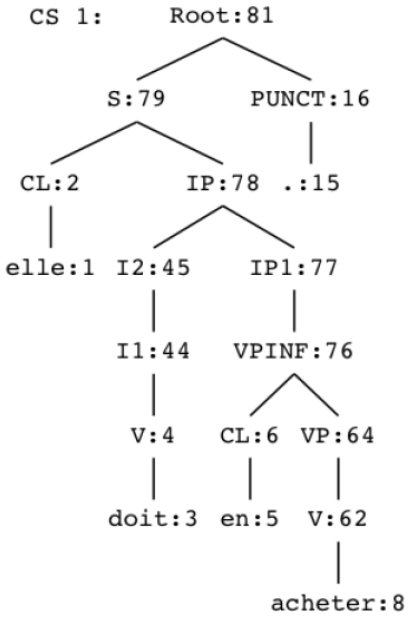


Figure 7 – *C-structure for example (8b).*

"Elle doit en acheter."

```
    [PRED    'DEVOIR<[1:PRO], [8:ACHETER]>'                          ]
15  SUBJ   1[PRED 'PRO'                                              ]
16         2[CASE NOM, GEND F, NUM SG, PERS 3                        ]
 3         8[PRED  'ACHETER<[1:PRO], [5:PRO]>'                       ]
 4         8[SUBJ  [1:PRO]                                           ]
44  XCOMP 62
45        64[OBJ   5[PRED 'PRO'                                      ]
78        76       6[CASE GEN, SPEC PART                            ]
79        77[AUX AVOIR, CFORM null, PASSIVE -, VFORM INF            ]
81  [CLAUSE_TYPE DECLAR, MOOD IND, PASSIVE -, TENSE PRES            ]
```

Figure 8 – *F-structure for example (8b).*

5.1. Obliques

Oblique (OBL) complements are typically realized by PPs and fall into two subclasses depending on whether the preposition is semantic or nonsemantic, the latter assigning case to a PP, cf. Butt *et al*. (1999), Bresnan (2001), etc. OBLs with the nonsemantic *de* are genitive marked and thus pronominalizable by *en*. **Figure 9-Figure 14** exemplify this analysis.

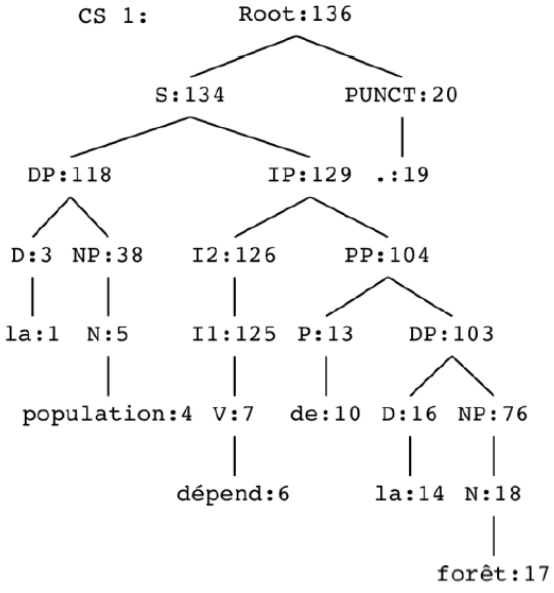


Figure 9 – *C-structure for example (1a).*

```
"La population dépend de la forêt."

      [PRED    'DÉPENDRE<[1:POPULATION], [10:FORÊT]>'                                    ]
             4[PRED 'POPULATION'                                                        ]
             5[CASE NOM, DCONTR -, DTYPE DEF, GEND F, MASS -, NUM SG, PERS 3, SPEC DEF  ]
      SUBJ  38
             1
             3
           118[                                                                         ]
            17[PRED 'FORÊT'                                                           ]
 19         18[CASE GEN, DCONTR -, DTYPE DEF, GEND F, NUM SG, PERS 3, PTYPE NOSEM, SPEC DEF
 20         76
  6         14
  7  OBL    16
125        103
126         10
129         13
134        104[                                                                      ]
136   [CLAUSE_TYPE DECLAR, MOOD IND, TENSE PRES                                       ]
```

Figure 10 – *F-structure for example (1a).*

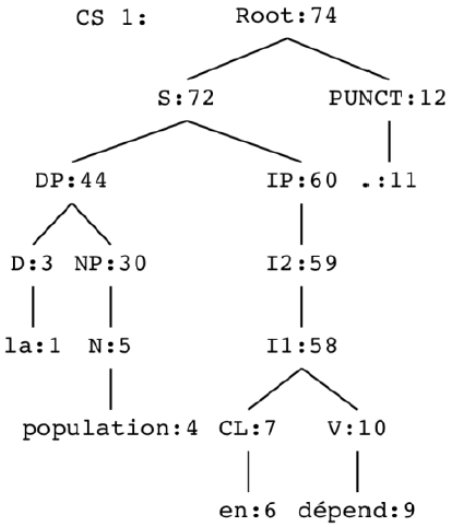


Figure 11 – *C-structure for example (1b).*

```
"La population en dépend."

      [PRED    'DÉPENDRE<[1:POPULATION], [6:PRO]>'                                      ]
             4[PRED 'POPULATION'                                                       ]
 11          5[CASE NOM, DCONTR -, DTYPE DEF, GEND F, MASS -, NUM SG, PERS 3, SPEC DEF
 12   SUBJ  30
  9          1
 10          3
 58         44[                                                                       ]
 59
 60   OBL    6[PRED 'PRO'                                                             ]
 72           7[CASE GEN                                                             ]
 74    [CLAUSE_TYPE DECLAR, MOOD IND, TENSE PRES                                      ]
```

Figure 12 – *F-structure for example (1b).*

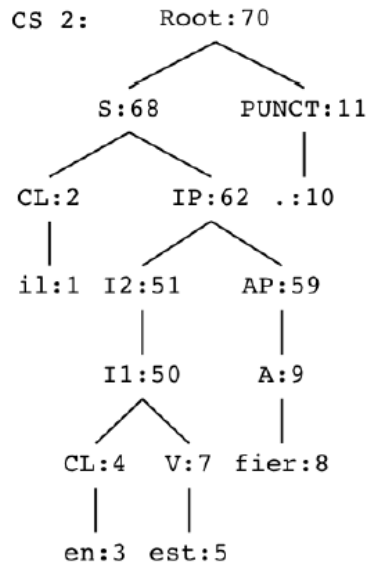


Figure 13 – *C-structure for example (3b).*

```
"Il en est fier."

    [PRED     'ÊTRE<[8:FIER]>[1:PRO]'                          ]
10  SUBJ   1 [PRED 'PRO'                                       ]
11         2 [CASE NOM, GEN MAS, NUM SG, PERS 3                ]
 5          [PRED  'FIER<[1:PRO], [3:PRO]>'                    ]
 7          [SUBJ  [1:PRO]                                     ]
50  XCOMP  8 [OBL    3[PRED 'PRO'                              ]
51         9 [       4[CASE GEN                                ]]
62          [                                                  ]
68       59 [ATYPE PREDIC                                      ]
70 [CLAUSE_TYPE DECLAR, EN +, MOOD IND, TENSE PRES             ]
```

Figure 14 – *F-structure for example (3b). The feature EN=+ licenses* en *in subject position (Section 5.4.).*

Entries for verbs and adjectives subcategorizing for a nonsemantic *de*-OBL follow the general pattern of (26), by substituting (↑OBL CASE)=c GEN for (↑OBL PCASE)=c DE. This constraint is satisfied by either *en* or a PP with the grammatical *de*. For generating examples with *de*-OBLs, we adapted Frank's c-structure rules in (40) to our genitive case analysis.

5.2. Adnominal adjuncts

In LFG, an ADJ(unct) is a non-subcategorized-for grammatical function (Bresnan, 2001). To generate examples with adnominal adjuncts headed by the nonsemantic *de*, we adapted Frank's NP-structure rules (47) and (48) to the DP-analysis, see provisional version in (54), to be revised in (59).

(54) DP → D NP
NP → …N… PP#0#1:{↓∈(↑ADJ)|(↑OBJ)=↓@(OT-MARK QDomain)};...

Rule (54) states that an NP consists of an N head optionally followed by a PP performing one of two functions: ADJ, see (6a) and (7a), or OBJ, see (4a). In the second disjunct, a call to the OT-MARK template defined by King (2004) ensures that the PP is preferably analyzed as an OBJ if the noun subcategorizes for one, as it is the case with nominal QForms (see next section). This reduces parsing ambiguity. **Figure 15-Figure 18** exemplify our analysis.

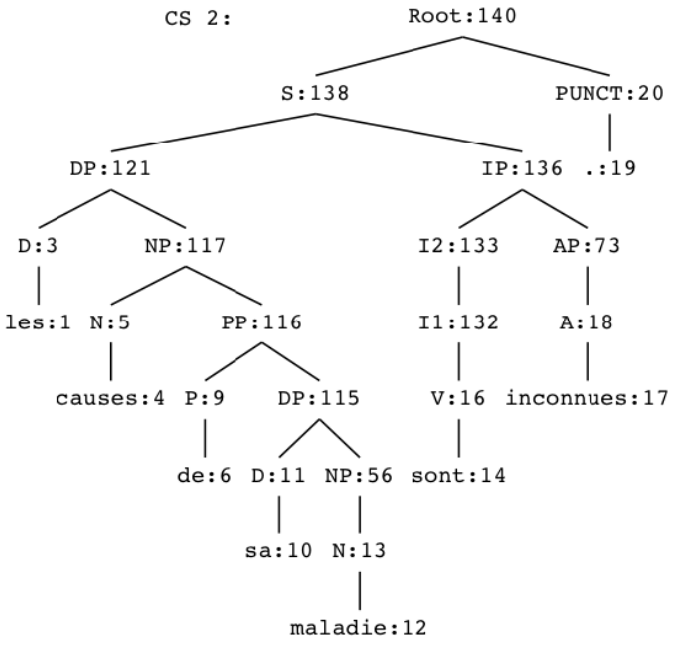


Figure 15 – *C-structure for example (6a).*

```
"Les causes de sa maladie sont inconnues."

     [PRED    'ÊTRE<[17:INCONNUE]>[1:CAUSE]'
          [PRED 'CAUSE'
                {  12[PRED 'MALADIE'
                      13
                      56POSS [PRED 'PRO'    ]
                      10     [NUM SG, PERS 3]
                      11CASE GEN, DCONTR -, GEND F, MASS -, NUM SG, PERS 3, PTYPE NOSEM, SPEC DEF
     SUBJ    4ADJ  {
             5      115
                     6
  19       117      9
  20         1    116[
  14         3
  16       121CASE NOM, DCONTR -, DTYPE DEF, GEND F, MASS -, NUM PL, PERS 3, SPEC DEF
 132
 133        17[PRED    'INCONNUE<[1:CAUSE]>'
 136XCOMP   18[SUBJ    [1:CAUSE]
 138        73[ATYPE PREDIC, POSITION POSTNOM]
 140CLAUSE_TYPE DECLAR, EN +, MOOD IND, TENSE PRES]
```

Figure 16 – *F-structure for example (6a).*

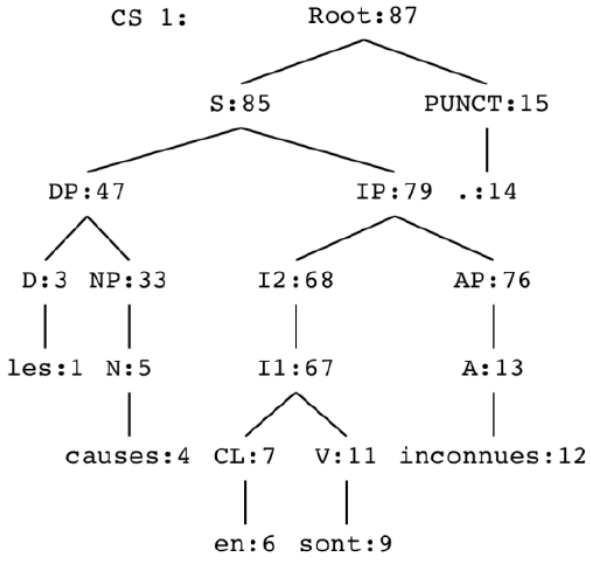


Figure 17 – *C-structure for example (6b).*

```
"Les causes en sont inconnues."

     [PRED    'ÊTRE<[12:INCONNUE]>[1:CAUSE]'
          4[PRED 'CAUSE'
          5
  14     33ADJ  {6[PRED 'PRO'  ]}
  15SUBJ    1   {7[CASE GEN   ]}
   9        3CASE NOM, DCONTR -, DTYPE DEF, GEND F, MASS -, NUM PL, PERS 3, SPEC DEF]
  11       47[
  67
  68       12[PRED    'INCONNUE<[1:CAUSE]>'
  79XCOMP  13[SUBJ    [1:CAUSE]
  85       76[ATYPE PREDIC, POSITION POSTNOM]
  87CLAUSE_TYPE DECLAR, EN +, MOOD IND, TENSE PRES]
```

Figure 18 – *F-structure for example (6b).*

5.3. Quantified terms

This subsection details the analysis of quantified terms. QForm classes were implemented by means of templates. However, to unburden the reader, we present full entries here.

Following Mittendorf and Sadler (2005)'s proposal for Welsh, we represent French quantified terms like (55)-(57) as numeral phrases (NumPs). Entries for simple cardinals follow the general pattern of (58); complex cardinals, e.g., (13), were not implemented.

(55) trois pommes
three apples

(56) ces trois pommes
these three apples

(57) trois de ces pommes
three DE these apples
'three of these apples'

(58) *trois*, Num
{(↑PRED)='TROIS<(↑OBJ)>'|(↑PRED)='TROIS' (↑CASE) ¬=ACC}
(↑NUM)=PL
@(DEFAULT (^ SPEC) INDEF)

The second line in (58) encodes subcategorization in form of a disjunction: the first disjunct states that the numeral requires an OBJ, as in (5) and (12), while the second allows for uses without an OBJ, which are restricted to non-accusative DPs (remember that "¬" represents negation). The third line specifies number. In French, all plural Num forms are underspecified for gender, only singular forms *un* and *une* manifest gender variation. The last line makes a call to the DEFAULT template (King, 2004), specifying that INDEF is the default value of SPEC. This can be overridden, e.g., by SPEC=DEM(onstrative), see (56). **Figure 19-Figure 26** exemplify this analysis.

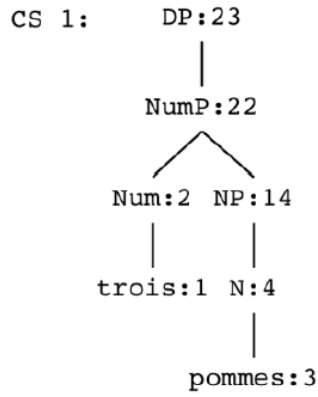


Figure 19 – *C-structure for example (55).*

"trois pommes"

```
   [PRED   'TROIS<[3:POMME]>'                          ]
 1 [      3[PRED 'POMME'                                ]
 2 [OBJ   4[CASE GEN, GEND F, MASS -, NUM PL            ]
22 [     14[                                           ]
23 [NUM PL, PERS 3, SPEC INDEF                         ]
```

Figure 20 – *F-structure for example (55).*

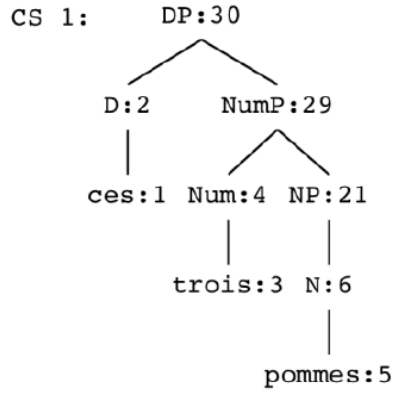


Figure 21 – *C-structure for example (56).*

```
"ces trois pommes"

  3 [PRED  'TROIS<[5:POMME]>'                                    ]
  4 [      5[PRED 'POMME'                                        ]]
 29 [OBJ   6[CASE GEN, GEND F, MASS -, NUM PL                    ]]
  1 [      21[                                                   ]]
  2 [                                                            ]
 30 [DCONTR -, NUM PL, PERS 3, SPEC DEM                          ]
```

Figure 22 – *F-structure for example (56).*

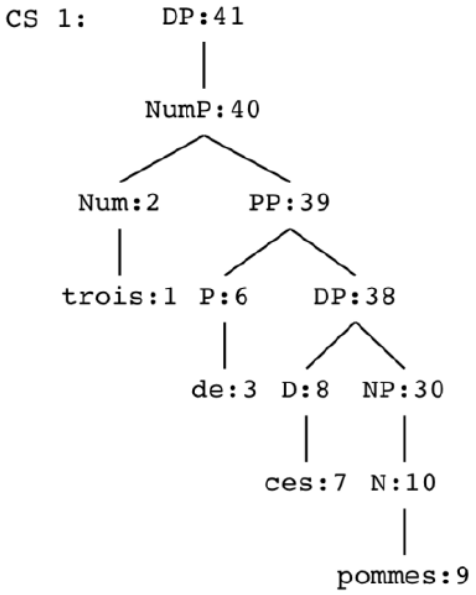


Figure 23 – *C-structure for example (57).*

```
"trois de ces pommes"

  [PRED  'TROIS<[3:POMME]>'                                                                  ]
  [     9[PRED 'POMME'                                                                       ]]
  [    10[CASE GEN, DCONTR -, GEND F, MASS -, NUM PL, PERS 3, PTYPE NOSEM, SPEC DEM         ]]
  [    30[                                                                                  ]]
  [     7[                                                                                  ]]
  [OBJ  8[                                                                                  ]]
  [    38[                                                                                  ]]
1 [     3[                                                                                  ]]
2 [     6[                                                                                  ]]
40[    39[                                                                                  ]]
41[NUM PL, PERS 3, SPEC INDEF                                                               ]
```

Figure 24 – *F-structure for example (57).*

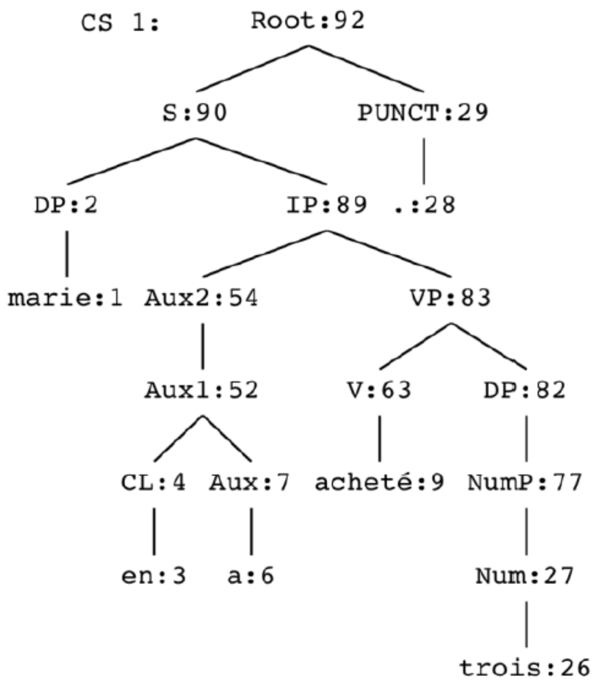


Figure 25 – *C-structure for example (5b).*

```
"Marie en a acheté trois."

28 PRED   'ACHETER<[1:MARIE], [26:TROIS]>'
29
 9 SUBJ  1 PRED 'MARIE'
         2 CASE NOM, DCONTR -, GEND F, NUM SG, PERS 3
63
83          PRED 'TROIS<[3:PRO]>'
         26
 6 OBJ   27 OBJ  3 PRED 'PRO'
 7       77      4 CASE GEN
52       82 CASE ACC, NUM PL, PERS 3, SPEC INDEF
54
89 CHECK [AVOIR +]
90 AUX AVOIR, CLAUSE_TYPE DECLAR, MOOD IND, PASSIVE -, TENSE COMPOUND_PAST, UNACC -, VFORM PART_PAST
92
```

Figure 26 – *F-structure for example (5b). CHECK features prevent overgeneration, but are not theoretically relevant (King, 2004).*

The attentive reader may have noticed that the OBJ is marked with genitive case both in **Figure 20** and **Figure 24**, although it is a bare NP in the former. The main motivation for this assumption is that both OBJs are pronominalizable by *en*, which we claim to be a pro-form for genitive-marked grammatical functions. Additional support comes from languages that mark the domain of a low-valued cardinal

in constructions analogous to **Figure 20** with a preposition (Welsh) or partitive (Finnish) or genitive case (Russian) (Corbett, 1978; Hurford, 2003). In French, this pattern is restricted to "nounier", higher-valued cardinals, see (13), but in Romanian *de* is required for cardinals from 20 upwards.

To generate nominals containing a NumP, we adapted Mittendorf and Sadler's DP-analysis of Welsh to the grammatical facts of French, see (59) and (60), a simplified version of the actually implemented rules.[12] The first rule states that a DP consists of an optional D followed by an NP or a NumP. The second rule states that a NumP consists of Num followed by either a PP or an NP functioning as OBJ, cf. (57) and (55). These two alternatives are represented in a disjunction. In the first disjunct, a constraining equation ensures that the preposition heading the PP be endowed with genitive case, i.e., it must be *de*. In the second disjunct, the second equation assigns genitive case to the OBJ and the last two handle head-complement agreement.

(59) DP → (D) {NP|NumP}
(60) NumP → Num {PP:(↑OBJ)=↓ (↓CASE)=c GEN|
NP:(↑OBJ)=↓ (↓CASE)=GEN (↑NUM)=(↓NUM) (↑GEND)=
(↓GEND)}#0#1

The remaining QForm classes of Table 2 are nouns, which we also assume to have a one-place and a zero-place variant, however, the former seems not to be constrained to non-OBJ positions, see (61), where the domain (seafood referred to in the previous sentence) is left unexpressed. The templates modeling the different classes include both variants, but, to simplify, we limit ourselves here to the monovalent variants.

(61) Nous avons acheté un kilo que nous avons dégusté […] (Google)
we have bought one kilo that we have enjoyed

12. The full rules handle, e.g., *****un un livre* 'an one book' vs. *les/ces/leurs deux livres* 'the/these/their two books' or **trois de livres* 'three of books'.

Collective numerals require the OBJ to be a plural count noun, see (62). The constraining equation in the last line enforces realization of the OBJ by a genitive-marked element, i.e., a PP with the nonsemantic *de* or the clitic *en*.

(62) *douzaine*, N
(↑PRED)='DOUZAINE(↑OBJ)'
(↑GEND)=F
(↑NUM)=SG
(↑OBJ MASS)=–
(↑OBJ NUM)=PL
(↑OBJ CASE)=c GEN

There is no need to impose any constraints on the number of the OBJ of fraction names, since they allow both mass nouns in singular and count nouns in singular and plural, see (4a) and (63). Accordingly, in entry (64), the number of the OBJ is underspecified.

(63) la moitié de la farine/des pommes
the half DE the:F.SG flour/DE.the.PL apples
'the half of the flour/of the apples'

(64) *moitié*, N
(↑PRED)='MOITIÉ(↑OBJ)'
(↑GEND)=F
(↑NUM)=SG
(↑OBJ CASE)=c GEN

Figure 27 and **Figure 28** illustrate the analysis of fraction names:

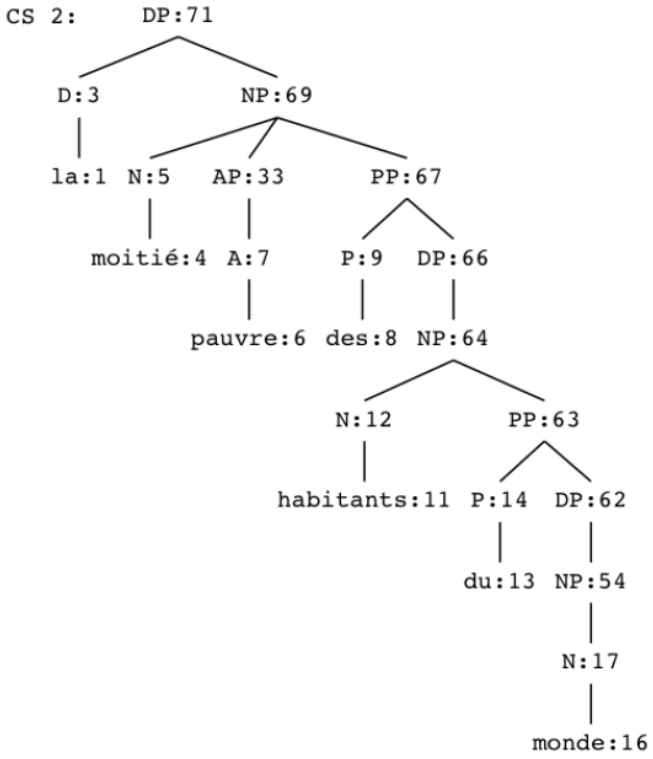


Figure 27 – *C-structure for example (16).*

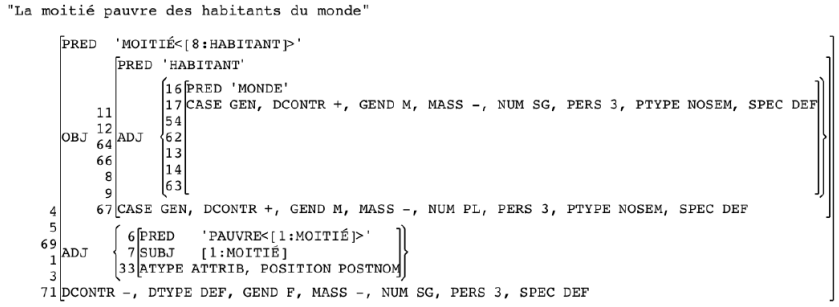


Figure 28 – *F-structure for example (16).*

Entries for measure names follow the same pattern of (64), see (14) and (65). Whether a count noun OBJ of a measure name must be plural, as suggested by Jones (1996, p. 219), is a question we leave for further research.

(65) un kilo de pommes
a kilo DE apples
'a kilo of apples'

5.4. Implementation of the climbed en

In this section, we detail the implementation of clitic climbing, which affects the genitive *en* and all other non-nominative clitics.

Figure 29 and **Figure 30** exemplify the treatment of climbing to non-causative full verb hosts, for which we reimplemented the corresponding c-structure rules proposed by Frank (1996).

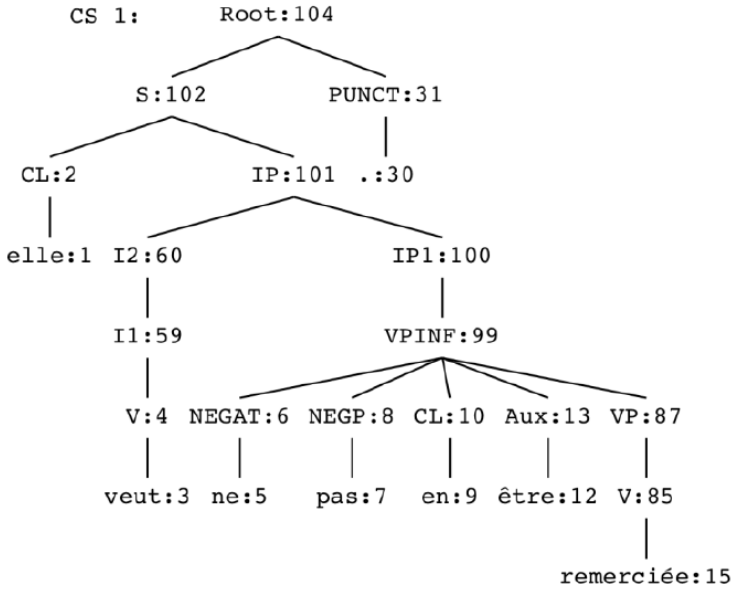


Figure 29 – *C-structure for example (39).*

```
"Elle veut ne pas en être remerciée."

      [PRED     'VOULOIR<[1:PRO], [5:REMERCIER]>'                                    ]
      [SUBJ      1[PRED 'PRO'                                                        ]
                 2[CASE NOM, GEND F, NUM SG, PERS 3]                                 ]
                [15[PRED  'REMERCIER<NULL, [1:PRO], [9:PRO]>'                        ]
                [85[SUBJ  [1:PRO]                                                    ]
                [87[      9[PRED 'PRO'                                               ]
  30            [12[OBL  10[CASE GEN                                                 ]
  31            [13[                                                                 ]
   3  XCOMP     [ 7[CHECK [PASS +_]                                                  ]
   4            [ 8[AUX AVOIR, CFORM null, EN +, NE +, NEG +, NEGP PAS, PASSIVE +, VFORM INF
  59            [ 5[                                                                 ]
  60            [ 6[                                                                 ]
 101            [99[                                                                 ]
 102           [100[                                                                 ]
 104 [CLAUSE_TYPE DECLAR, MOOD IND, PASSIVE -, TENSE PRES                            ]
```

Figure 30 – *F-structure for example (39).*

Our implementation of causative constructions, however, is an adaptation of Yates (2002)'s biclausal analysis, because Frank's analysis does not cover examples such as (66).

(66) La pauvreté en=a=fait dépendre la population.
the poverty EN=have=made depend.INF the population
'Poverty made the population depend on it.'

Figure 31-Figure 36 illustrate the parsing of different structures with *en* and *faire* 'make'.

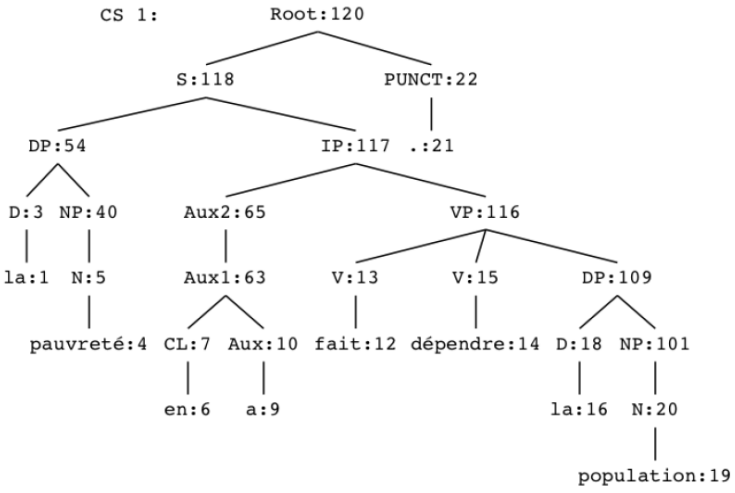


Figure 31 – *C-structure for example (66).*

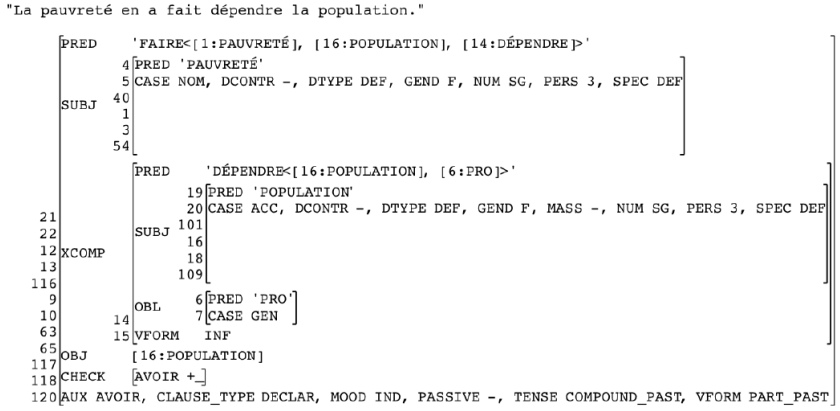


Figure 32 – *F-structure for example (66).*

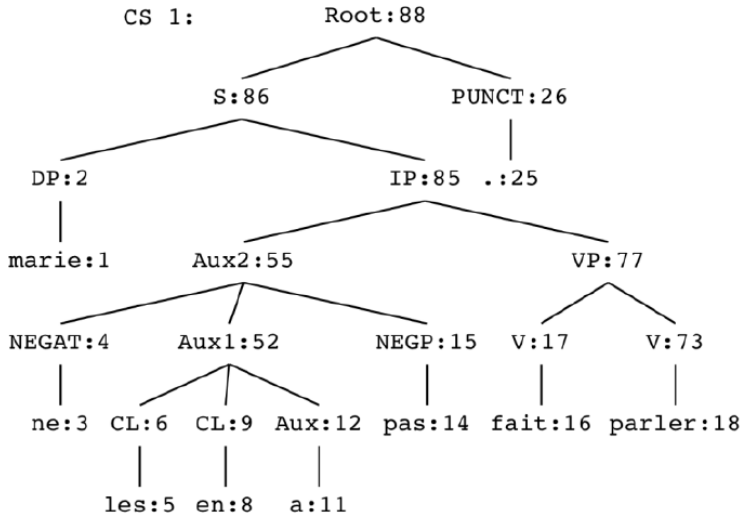


Figure 33 – *C-structure for example (37).*

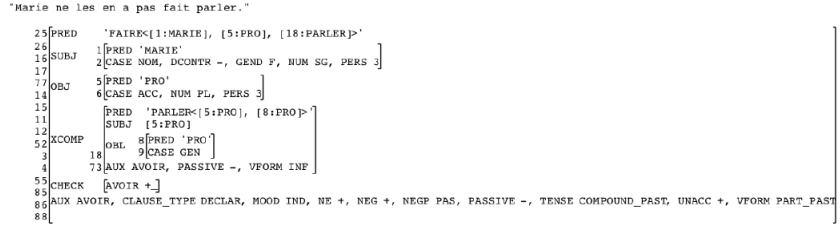


Figure 34 – *F-structure for example (37).*

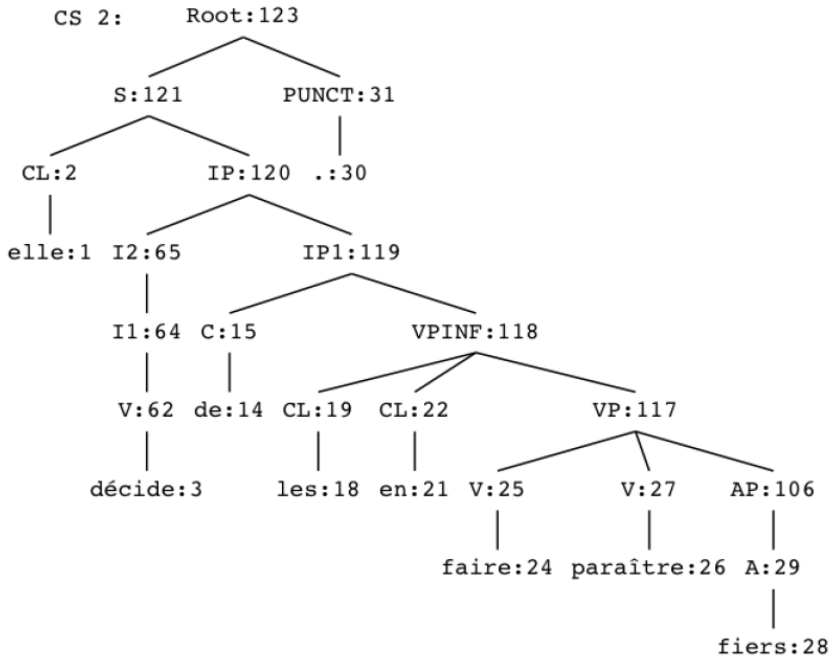


Figure 35 – *C-structure for example (10).*

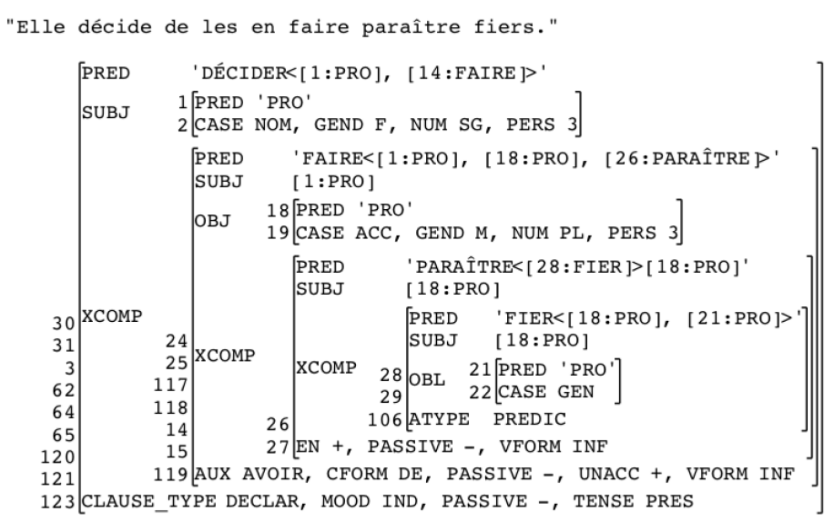


Figure 36 – *F-structure for example (10).*

In case of auxiliary constructions, we also departed from Frank's approach. Instead of treating auxiliaries as verbs subcategorizing for a VCOMP, we preferred to implement them as items deprived of a PRED attribute. As such, they do not govern any grammatical function, only contributing to the f-structure of the sentence with morphosyntactic features like person, number, tense, etc. For the sake of computational efficiency, however, auxiliaries were assigned a special c-structure category Aux, instead of being differentiated from full verbs by means of features, which would demand higher processing costs (Butt *et al*., 1999). Analogously to Frank's I1 and I2 categories, see (40), the complexes formed by auxiliary, pronominal, and negation clitics were labeled Aux1 and Aux2, see **Figure 31** and **Figure 35**.

The functional annotations for pronominal clitics follow the general pattern of (45), however, they were adapted to our analysis of *en* and *de*, as formulated in (51) and (52), and extended to cover all constructions (1)-(8).

In our implementation, we made extensive use of metacategories, a powerful resource of XLE that is not available in the formalism in which Frank's grammar was implemented. A metacategory is a variable for one or more c-structure non-terminal nodes with the respective annotations. A metacategory can be non-recursively used in the definition of another metacategory. Templates can also be used in these definitions. Analogously to macros in programming languages, metacategories, like templates, allow for code abstraction, enhancing readability and maintainability of a computational grammar. Thanks to metacategories, it was not necessary to stipulate different positional variants for *en* and the other pronominal clitics, as Frank does.

Generalizing (45), restricted to OBLs with the nonsemantic *de*, we defined the template (67), which can be used with any governed grammatical function F given as parameter, see (69)-(71). The constraint on clitic climbing excluding an XCOMP with a complementizer form (CFORM) is encapsulated in template (68). In (69) and (70), we make use of template (67) to define metacategories CL-IO and CL-OBL for dative indirect object (OBJ2) and genitive OBL clitics, respectively, see (72).

(67) CL-GF(F)={(↑F)=↓|(↑XCOMP+ F)=↓ @NO_CFORM}.
(68) NO_CFORM=¬(↑XCOMP+ CFORM).
(69) CL-IO=CL: @(CL-GF OBJ2) @DAT.
(70) CL-OBL=CL: @(CL-GF OBL) @GEN.
(71) DO-CL(C)=@(CL-GF OBJ) C.
(72) Marie lui=en=a donné.
Mary him.DAT=EN=has given
'Mary gave him some.'

An OBJ clitic can bear either genitive or accusative case, depending on whether it is a partitive object or not, so we defined template (71) for clitics performing this function, where parameter C is the case of the clitic. This template, in turn, is used to define templates (73) and (74) for accusative-marked and genitive-marked clitic OBJs, respectively. The latter is assigned SPEC=PART, so that it coheres in specification with partitive DP-objects, see **Figure 6** and **Figure 8**. The different case markings of clitics are encoded by means of templates analogous to (75).

(73) ACC-DO-CL=@(DO-CL @ACC)...
(74) GEN-DO-CL=@(DO-CL @GEN) (↓SPEC)=PART.
(75) ACC=(↓CASE)=ACC.

The constraints on *en* as ADJ(unct) to a clausal SUBJ or OBJ, see (6b) and (7b), are encoded in (76) by means of a disjunction. The constraint equation in the first disjunct requires feature EN to have a positive value. This is provided by verbs enabling *en* in subject position, e.g., *être* 'be', whose entries have (↑EN)=+.

(76) CL-ADJ = CL:{↓∈(↑SUBJ ADJ) (↑EN)=c+|↓∈(↑OBJ ADJ)} @GEN.

Analogously, to capture the use of *en* as OBJ of a QForm in OBJ or SUBJ position, cf. (4b), (5b), and (18), the metacategory CL-QFORM-DOM is defined following the general pattern in (67).

All functions of *en* are collapsed into the metacategory (77), which is used in (78), together with metacategories for the other clitic types, to disjunctively represent all possible clusters of genitive and third-person accusative and dative clitics.

(77) CL-GEN={CL-DO-GEN|CL-OBL|CL-ADJ|CL-QFORM-DOM}.
(78) CL-PRON={CL-DO-ACC CL-IO (CL-GEN)|CL-DO-ACC (CL-GEN)|CL-IO (CL-GEN)|CL-GEN}.

This, in turn, enables us to formalize in a single rule the optional attachment of a varied range of pronominal clitics to a verbal head, as shown in (79).

(79) I1 → (CL-PRON) V.

6. Implementation methodology and evaluation

An LFG grammar is a declarative model of a language fragment, encoding constraints at different levels. Due to the complexity of these constraints, the implementation of a non-trivial computational grammar fragment must be an incremental process. One starts with a very small grammar capable of analyzing simple examples and progressively extends it to cover an increasingly larger subset of the phenomena to be modeled. These successively more complex fragments must be tested not only on grammatical sentences, but also on examples that violate the postulated constraints. These two data types are labeled positive test set and negative test set. Thanks to the declarative nature of the formalism, an LFG grammar can be used for both analysis and generation. These two dimensions can be evaluated by the positive test set and the negative test set, respectively.[13]

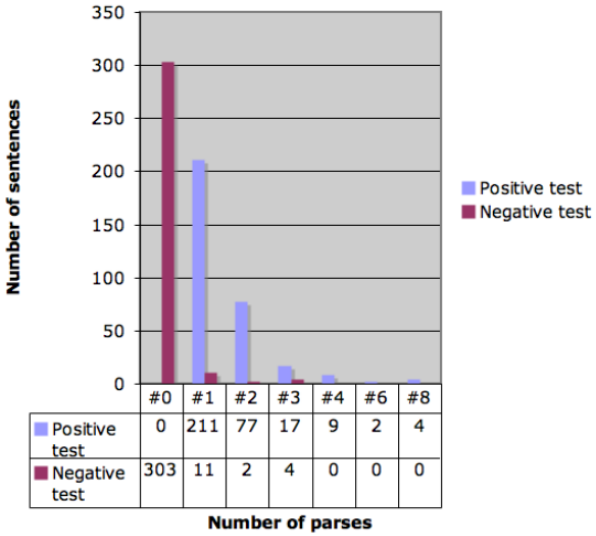
In our case, we did not have to start the grammar development from scratch. Two previous works were available to start from. On the one hand, we could reuse large portions of code from FrGramm, which covers basic French syntax phenomena, although it cannot handle non-subject pronominal clitics, partitive DPs, and quantified expressions. On the other hand, Frank's grammar already handles the relationship between *de* and *en* in (1) and (3) and can analyze constructions (2a), (4a) and (6a)-(8a), so we could reimplement the corresponding c-structure rules and annotations in XLE and extend them to cover the other constructions. The implementation of causative *faire*, however,

13. On the development and testing of LFG grammars with XLE, see Butt *et al.* (1999).

DELTA
37.1
2021

demanded an extra effort, since Frank (1996) only handles a small subset of these constructions.

The final grammar fragment was tested on a positive test set with 320 grammatical sentences and on a negative test set with an equal number of ungrammatical examples. **Figure 37** presents the results.



	#0	#1	#2	#3	#4	#6	#8
Positive test	0	211	77	17	9	2	4
Negative test	303	11	2	4	0	0	0

Figure 37 – *Parsing results for the positive test set and the negative test set.*

The positive test includes all grammatical examples of this paper, except for (13) and (30).[14] As **Figure 37** shows, all sentences received at least one parse, i.e., a valid f-structure according to the grammar. 65,9% received exactly one parse and 24.1% exactly two. Only 10% of the sentences were assigned between 3 and 8 parses. Thus, the grammar is lowly ambiguous, which is a desirable feature from a natural language perspective. In **Figure 38**, we have the parsing results for three sentences from the positive test set treated as ambiguous by the parser.

14. Grammar and test sets will be made freely available on the FrGramm repository at https://github.com/lfg-french-grammar/.

```
# EXAMPLE: (6a)¶
¶
Les causes de sa maladie sont inconnues.  (2 0.040 28)¶
¶
# EXAMPLE: (10)¶
¶
Elle décide de les en faire paraître fiers.  (2 0.020 55)¶
¶
# EXAMPLE: (80)¶
¶
Le sac en est plein.  (3 0.010 25)¶
¶
```

Figure 38 – *Parsing results generated by XLE for a sample from the positive test set. The first number inside brackets after each parsed example indicates the total of valid f-structures assigned.*

Ambiguity arises lexically or structurally. The preposition *de* exemplifies the first type. For example, (6a) is assigned two valid f-structures, although only the one with the nonsemantic *de* seems plausible. The second type results from the functional ambiguity of *en*, which can realize either a complement or an adjunct, both in object and subject position. Since many complements are facultative, sentences such as (80) are assigned two valid f-structures in addition to the one where *en* functions as complement of adjective *plein* 'full'. In the other two less preferred f-structures, *en* functions as a complement and an adjunct of the QForm *sac* 'bag', respectively. Likewise, (10) receives an additional parse where *en* is an adjunct to the clitic object. Such reading is clearly spurious, because pronominal clitics cannot be modified by adjuncts, showing the need to further constrain the annotations of (76).

(80) Le sac en=est plein.
the bag EN=is full
'The bag is full of it.'

The negative test set includes all invalid constructions exemplified in this paper. It was built by systematically injecting errors into sentences of the positive test set. For example, (82) derives from (81) by pluralizing the noun, thus getting the determiner-noun agreement wrong, cf. (33).

(81) Elle=achète un livre.
she=buys one:M.SG book(M):SG
'She buys one book.'
(82) *Elle=achète un livres.
she=buys one:M.SG books(M):PL

Similarly, we obtained (83) from (8b) by left-adjoining the clitic to the modal verb, thereby violating the prohibition on clitic climbing imposed by this verb.

(83) *Elle=en=doit acheter.
she=EN=must buy

Figure 39 shows the parsing results for analogous ungrammatical variants of example (38), all of which violate ordering constraints between clitics, negation, and modal verb.

```
# SENTENCE_ID: 081

Elle ne en le veut pas faire parler. (0 0.000 7)

# SENTENCE_ID: 082

Elle ne l'en pas veut faire parler. (0 0.010 0)

# SENTENCE_ID: 083

Elle ne l'en veut pas faire parler. (0 0.000 16)

# SENTENCE_ID: 084

Elle ne pas en le veut faire parler. (0 0.010 0)

# SENTENCE_ID: 085

Elle ne pas l'en veut faire parler. (0 0.000 0)

# SENTENCE_ID: 086

Elle ne veut pas en le faire parler. (0 0.010 13)
```

Figure 39 – *Parsing results generated by XLE for ungrammatical variants of example (38).*

The negative test results show the grammar is highly constrained, since it blocks 94,7% of the ungrammatical examples. One of the 17 false positives is (84). This example is assigned a valid f-structure because *en* is analyzed as an adjunct of the subject clitic, to which e*n* cannot cliticize. There are 6 other similar examples, showing the need to fine-tune the metacategory definition in (76). The other 10 false positives involve the structural ambiguity of *en* and/or the lexical ambiguity of prepositions, as in (85) and (86). In the former example, *en* is an adjunct of *chamber* 'room', which does not seem to make sense. The latter example is assigned a valid (though nonsensical) f-structure where *de* is a semantic preposition introducing an adjunct.

(84) *Il=en=est fier de la victoire.
he=EN=is proud DE the:F.SG victory
(85) *La chambre en=est pleine de livres.
the room EN=is full DE books
(86) *La dame visite de reines vaillantes.
the lady visits DE queens valiant:F.PL

Summing up, the grammar can be said to be empirically valid, inasmuch as it was tested on a large amount of data. On the one hand, it analyses all 320 grammatical examples and, on the other hand, blocks most of the 320 ungrammatical examples. As a fragment, however, the grammar probably still has gaps that testing on more data could reveal.

7. Conclusion

We reported on an LFG implementation of French *en* and *de* in a wide range of constructions. Previous LFG approaches only cover a small subset of these structures and handle their heterogeneity by means of lexical ambiguity. Instead, we proposed a single variant for each of the involved elements, whose entries are linked by the genitive case.

Our proposal is a lexicalist reformulation of the main insight behind Kayne (1975)'s analysis, namely that *de* and *en* represent a single abstract category. There are, however, two important differences. For Kayne, *en* is derived from a deep structure with *de* by means of syntactic transformations, which are problematic form the parsing

perspective. By contrast, we claim that both *en* and *de* map to a genitive feature in f-structure, dispensing with any transformations. The second difference is that our grammar licenses a QForm in preverbal subject position governing an *en* OBJ, see (18). This construction does occur in real texts, so that a parser should be able to analyze it. However, Kayne (1975) considers it ungrammatical, in which he was followed by the subsequent literature.

The other reviewed approaches abandoned the pursuit of a unified treatment of *de* and *en*. We should point out other important features that set our proposal apart. First, we correctly handle both number and gender agreement in constructions like (34), while Hulk limits herself to the former. Second, differently from Carlier *et al.* (2003), *en* can represent outside the verbal domain not only noun complements, but also adjuncts to nouns and complements of adjectives and cardinal numbers.

The implementation in XLE enabled the grammar to be extensively tested on a large amount of examples. The results revealed a high level of coverage and low overgeneration: all grammatical sentences were successfully analyzed, with a low ambiguity rate, while only 5,3% of the ungrammatical examples were assigned valid f-structures.

As opportunities for further research, we suggest: implementing other QForm types; investigating the occurrence of *en* in preverbal subject position in large corpora and modeling the constraints it is subject to; and reducing ambiguity and overgeneration of the grammar.

References

ALENCAR, Leonel Figueiredo de. 2017. A computational implementation of periphrastic verb constructions in French. *Alfa: Revista de Linguística*, 61, pp. 437-466.

BRESNAN, Joan. 2001. *Lexical-functional syntax*. Malden: Blackwell.

BUTT, Miriam *et al*. 1999. *A grammar writer's cookbook*. Stanford: CSLI.

CARLIER, Anne; GOYENS, Michèle; LAMIROY, Béatrice. 2013. DE: A genitive marker in French? Its grammaticalization path from Latin to French. In: Carlier, Anne; Vanderstraete, Jean-Christophe. (eds.). *The genitive*, pp. 141-216. Amsterdam: Benjamins.

CORBETT, Greville G. 1978. Universals in the syntax of cardinal numerals. *Lingua*, 46: pp. 355-368.

FÁBREGAS, Antonio. 2017. Adjectival and genitival modification. In: Dufter, Andreas; Stark, Elisabeth. *Manual of Romance morphosyntax and syntax*, pp. 771-803. Berlin: De Gruyter.

FRANK, Anette. 1996. Eine LFG-Grammatik des Französischen. In: Berman, Judith; Frank, Anette. *Deutsche und französische Syntax im Formalismus der LFG*, pp. 97-244. Tübingen: Niemeyer.

GRIMSHAW, Jane. 1982. On the lexical representation of Romance reflexive clitics. In: Bresnan, Joan (ed.). *The mental representation of grammatical relations*, pp. 87-148. Cambridge: MIT Press.

HEAP, David; Oliviéri, Michèle; Palasis, Katerina. 2017. Clitic pronouns. In: Dufter, Andreas; Stark, Elisabeth (eds.). *Manual of Romance morphosyntax and syntax*, pp. 183-229. Berlin: De Gruyter.

HULK, Aafke. 1983. La syntaxe du pronom *en* dans la construction quantitative. *Revue québécoise de linguistique*, 13: pp. 167-199.

HURFORD, James R. 2003. The interaction between numerals and nouns. In: Plank, Frans (ed.). *Noun phrase structure in the languages of Europe*, pp. 561-620. The Hague: Mouton de Gruyter.

JONES, Michael Allan. 1996. *Foundations of French syntax*. Cambridge: Cambridge University Press.

KAYNE, Richard. 1975. *French syntax: The transformational cycle*. Cambridge: MIT Press.

KING, Tracy Holloway. 2004. *Starting a ParGram grammar*. http://ling.uni-konstanz.de/pages/xle/doc/PargramStarterGrammar/starternotes.html

KLENK, Ursula. 2003. *Generative Syntax*. Tübingen: Narr.

LAGAE, Véronique. 1997. *En* quantitatif: Pronom lié à la function object ou à une position? *Travaux de Linguistique*, 35, pp. 103-114.

MILNER, Jean-Claude. 1978. *De la syntaxe à l'interprétation: Quantités, insultes, exclamations*. Paris: Le Seuil.

MITTENDORF, Ingo; SADLER, Louisa. 2005. Numerals, nouns and number in Welsh NPs. In: Butt, Miriam; King, Tracy H. (Eds.). *Proceedings of the LFG05 Conference.* Stanford: CSLI. https://web.stanford.edu/group/cslipublications/cslipublications/LFG/10/pdfs/lfg05mittendorfsadler.pdf.

MÜLLER, Stefan. 2018. *Grammatical theory: From transformational grammar to constraint-based approaches* (2nd ed., Vol. 1). Berlin: Language Science Press.

SCHWARZE, Christoph. 1996. Die farblosen Präpositionen des Französischen: vage Prädikate oder Kasusmarker? *Romanische Forschungen*, 108, pp. 1-22.

_____. 2001. On the representation of French and Italian clitics. In: Butt, Miriam; King, Tracy Holloway (Eds.). *Proceedings of the LFG01 Conference*. Stanford: CSLI. http://web.stanford.edu/group/cslipublications/cslipublications/LFG/6/pdfs/lfg01.pdf.

_____. 2012. Romance clitic pronouns in lexical paradigms. In: Sascha Gaglia and Marc-Olivier Hinzelin (Eds.). *Inflection and Word Formation in Romance Languages*. Amsterdam – Philadelphia: John Benjamins. 119-140.

SCHWARZE, Christoph; ALENCAR, Leonel Figueiredo de. 2016. *Lexikalisch-funktionale Grammatik: Eine Einführung am Beispiel des Französischen mit computerlinguistischer Implementierung*. Tübingen: Stauffenburg.

YATES, Nicholas. 2002. French causatives: A biclausal account in LFG. In: Butt, Miriam; King, Tracy Holloway (Eds.). *Proceedings of the LFG02 Conference*. Stanford: CSLI. https://web.stanford.edu/group/cslipublications/cslipublications/LFG/7/pdfs/lfg02yates.pdf.

Recebido em: 06/09/2019
Aprovado em: 09/08/2020